Deploying Natural Language Processing for Social Science Analysis

Karin Verspoor, Los Alamos National Laboratory

Antonio Sanfilippo, Pacific Northwest National Laboratory

Mark Elmore, Oak Ridge National Laboratory

Ed Mackerrow, Los Alamos National Laboratory

Content Analysis for communication research

- Content analysis:
  - methodology for systematic analysis of characteristics of messages
  - used by social scientists to investigate the nature of communications in terms of quantitative variables from which inferences can be drawn about context, meaning, and/or intents
  - Examples (from Neuendorf, 2002):
    - Prevalence of violence in top-grossing films of the 1990s
    - Importance of physical attractiveness in personal ads in newspapers
    - Analysis of the language of schizophrenics
- Harold Lasswell on communication:

Who says *what*, to *whom*, *why*, to *what extent*, and with *what effect*?

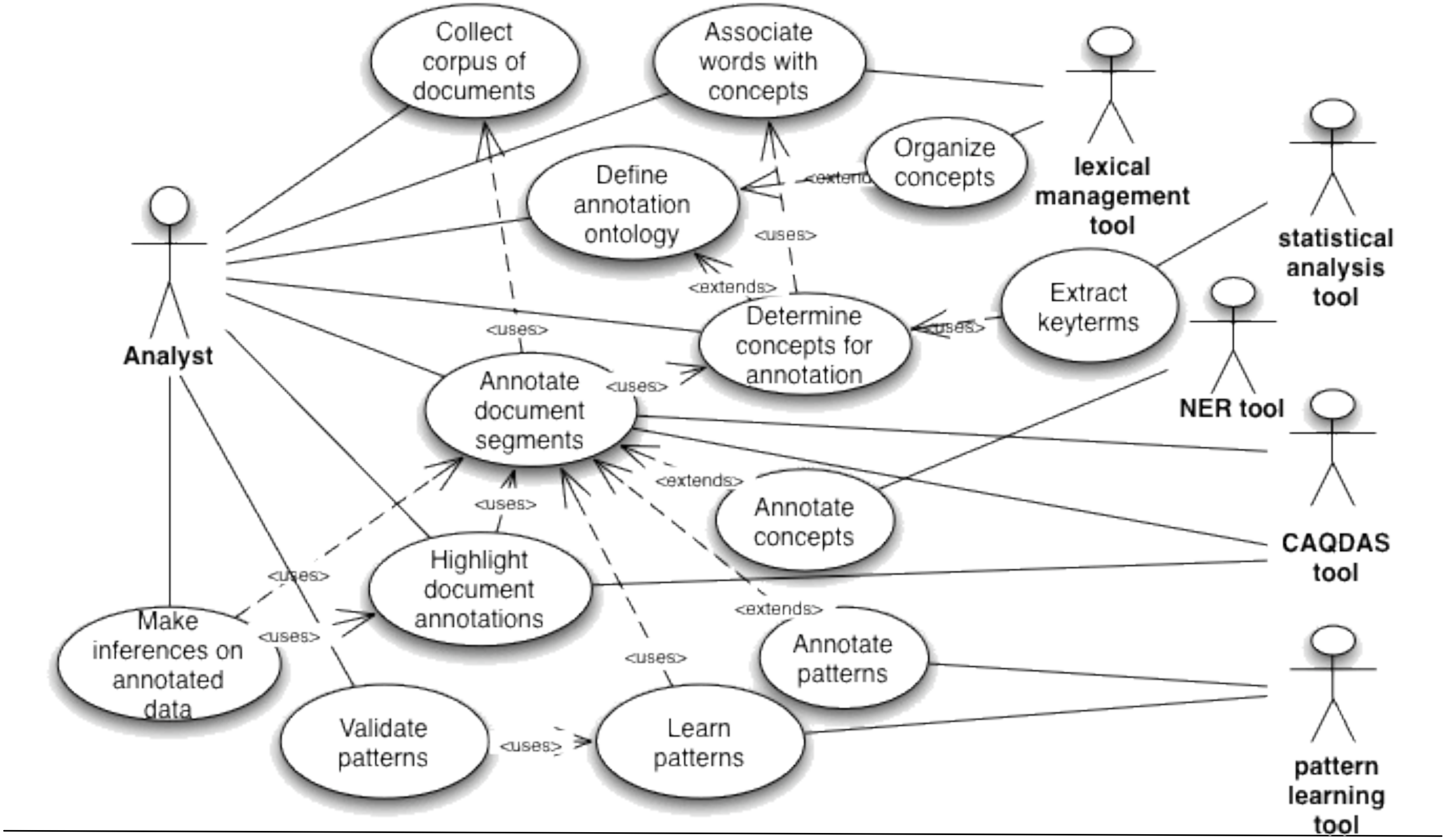
Methods for content analysis

- Fundamentally quantitative in nature
  - Neuendorf: “A content analysis has as its goals a numerically based summary of a chosen message set.”
  - counts of words or phrases in a given dictionary/category set
  - measurements of the amounts of variables, as determined from *codes* annotated on *messages*
    - This can be meta-data, e.g. date, duration, role or age of a character in a story, socio-economic status of a speaker, religious affiliation
    - Some of these attributes may be directly identifiable in the message
- CAQDAS: Computer-Assisted Qualitative Data Analysis Software
  - Software tools that support dictionary-based and manual coding
  - Frequency output
  - Searching for boolean co-occurrences of codes
  - Much coding, in particular of more complex patterns, must still occur manually

Natural Language Processing for content analysis

- Natural Language Processing (NLP) aims to exploit linguistic analysis techniques for text to determine
  - *syntactic* (structural) properties
  - *semantic* (meaning) properties
  - *statistical* properties
- Augmentation of content analysis tools with NLP techniques can facilitate higher-throughput content analysis and ultimately more automation of content analysis
  - named entity recognition
  - word-sense disambiguation
- Insights from the NLP research community can further enhance computer-assisted content analysis
  - organizing concepts (lexical management, ontologies)
  - learning information extraction patterns

Use case for NLP-augmented content analysis

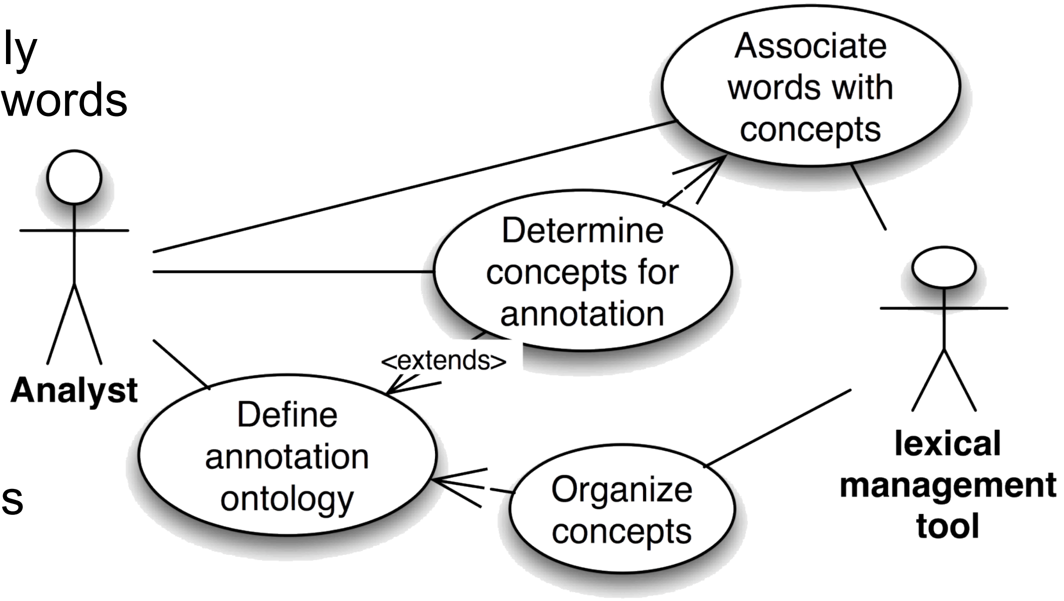


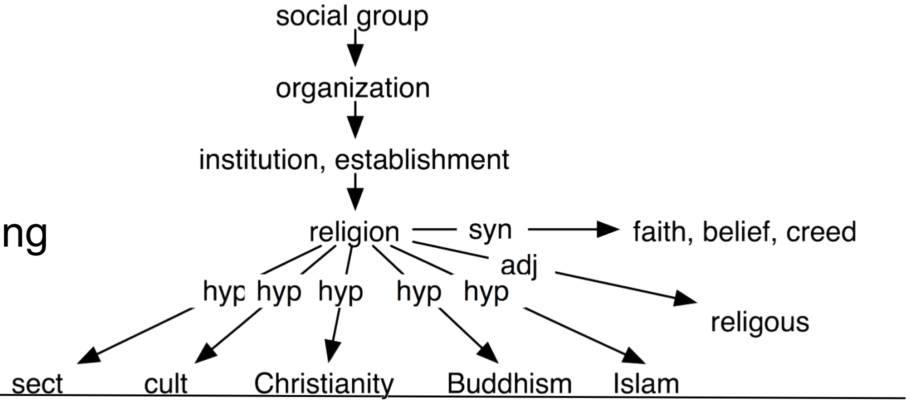
Augment CAQDAS with NLP

- Integrate NLP tools into the CAQDAS environment
  - Retain manual coding and boolean search capabilities: final decisions rest with the content analyst
  - Bring in more sophisticated text analysis tools to
    - facilitate message vocabulary assessment
    - automatically code many core semantic categories
- Utilize coded messages as a training set for further automation
  - generalize from examples using linguistic properties
  - identify patterns that correspond to more complex (e.g. relational) codes

Lexical management for content analysis

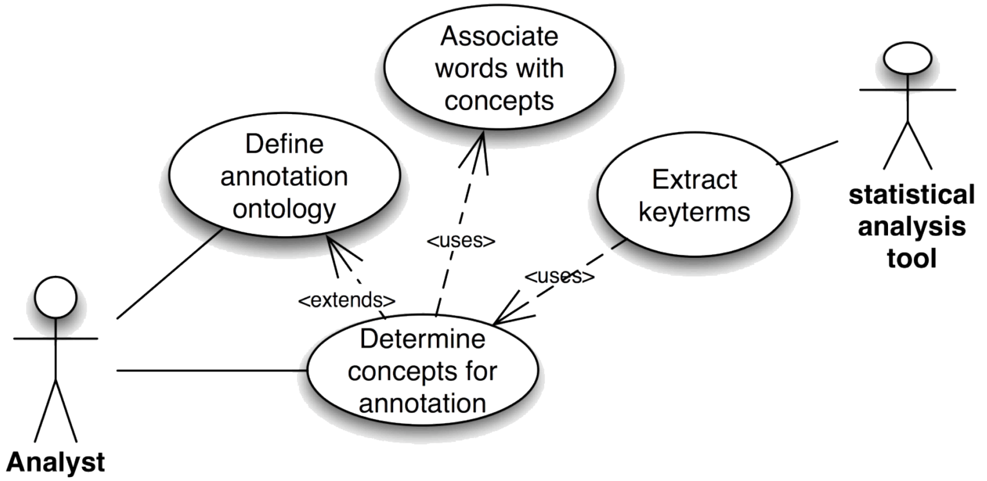
- Beyond dictionaries: hierarchically organized semantic networks of words
  - Synonyms
  - Hyper/hyponyms
- for concept refinement or generalization
- enabling inference over codes
- manage surface variants of terms “behind the scenes”
  - singular vs. plural
  - verb inflections
- Application-specific annotation ontology
  - Provides a controlled vocabulary
  - Ensures greater consistency in coding


Associate words with concepts
Determine concepts for annotation
<extends>
Analyst
Define annotation ontology
Organize concepts
lexical management tool



Statistical analysis for keyterm extraction

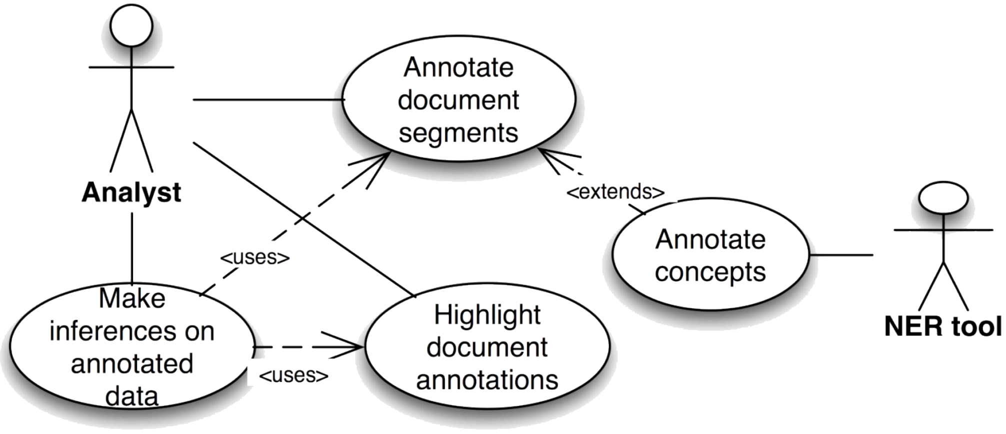
- Highly frequent words in a text are not always meaningful (cf. function words, “stop” words)
- Statistical tests can help to hone in on the real content-bearing terms, both words and phrases
  - TF.IDF for words
  - Chi-squared test of independence of words in a bigram


Associate words with concepts
Define annotation ontology
Extract keyterms
statistical analysis tool
<uses>
<uses>
<extends>
Determine concepts for annotation
Analyst

- Statistically derived keyterms can form the foundation of the annotation ontology for the topic of the content analysis

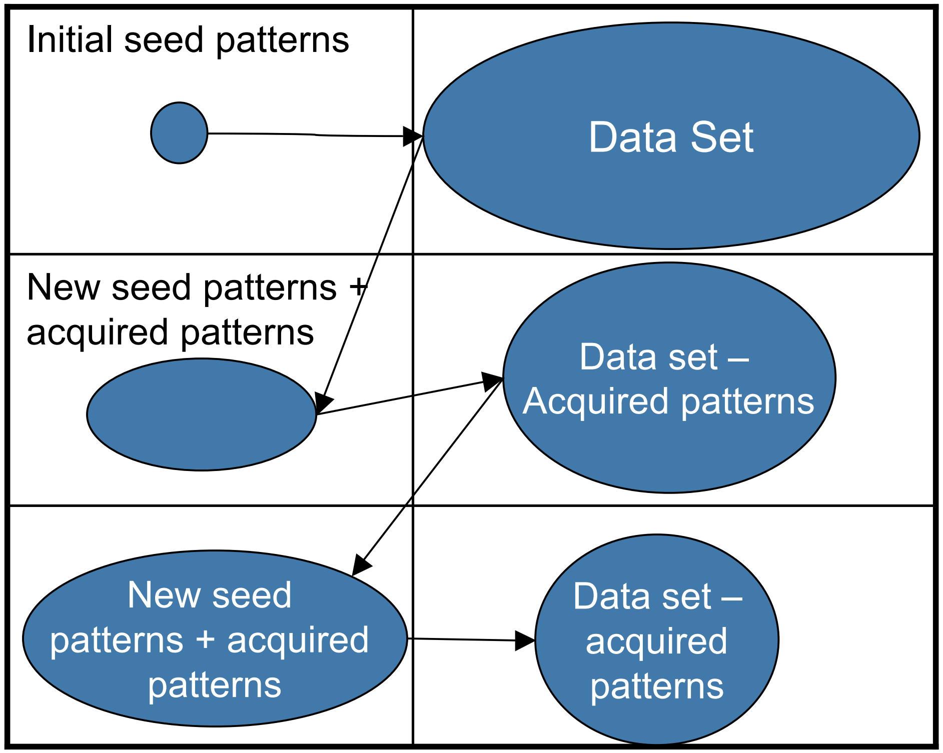
Named Entity Recognition (NER): *Who* says What to *Whom*?

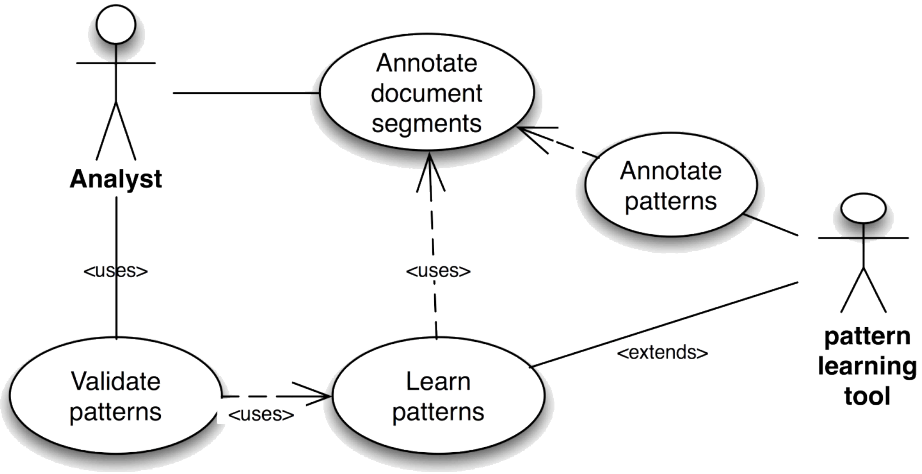
- NER tools aim to *automatically* identify and label certain basic semantic categories of terms in natural language text
  - People
  - Organizations
  - Places
  - Things (e.g. diseases, weapons, etc.)
- These will likely be useful categories to start coding from
- Allows the computational system to make a first pass at coding; freeing the human coder to identify more complex codes


Analyst
Annotate document segments
<extends>
Annotate concepts
NER tool
<uses>
Make inferences on annotated data
<uses>
Highlight document annotations

- Using supervised learning and sufficient training data, named entity recognizers can be trained to recognize other categories of terms relevant to the content analysis domain

Pattern Learning


Initial seed patterns
Data Set
New seed patterns + acquired patterns
Data set – Acquired patterns
New seed patterns + acquired patterns
Data set – acquired patterns



- Use annotations in coded data subset as seed patterns
- Use bootstrap semi-supervised learning algorithm to find best matches with seed patterns in large data set
  - Thelen & Riloff, 2002
  - Using linguistic context of codes

Case Study: Extraction of Social Movement Theory Signatures

- Annotation study undertaken at Pacific Northwest National Laboratory
  - *Antonio Sanfilippo, Andrew Cowell, Annie Boek and Stephen Tratz*
- Identify and code relevant text segments in harvested documents
  - Create ground truth data set semi-manually
  - Utilize NLP techniques to facilitate manual coding
  - Validate manual coding through inter-coder agreement
- Construct queries to find signatures of Social Movement Theory (e.g. Political Constraints and Frames) and quantify their strength

- **Political constraint**
  - **Source** = Egyptian Regime
  - **Policy** = Repression
  - **Target** = Sinai Bedouins
- **Strength** = Supported by 20% of the data consulted
- **Provenance** = [list document sources]

- **Diagnostic anti-system Frame**
  - **Promoter** = Muslim Brotherhood
  - **Intent** = Denounce
  - **Issue** = Repression
  - **Target** = Egyptian Regime
  - **Audience** = Egyptian Citizens
- **Strength =** Supported by 75% of the data consulted
- **Provenance** = [list document sources]

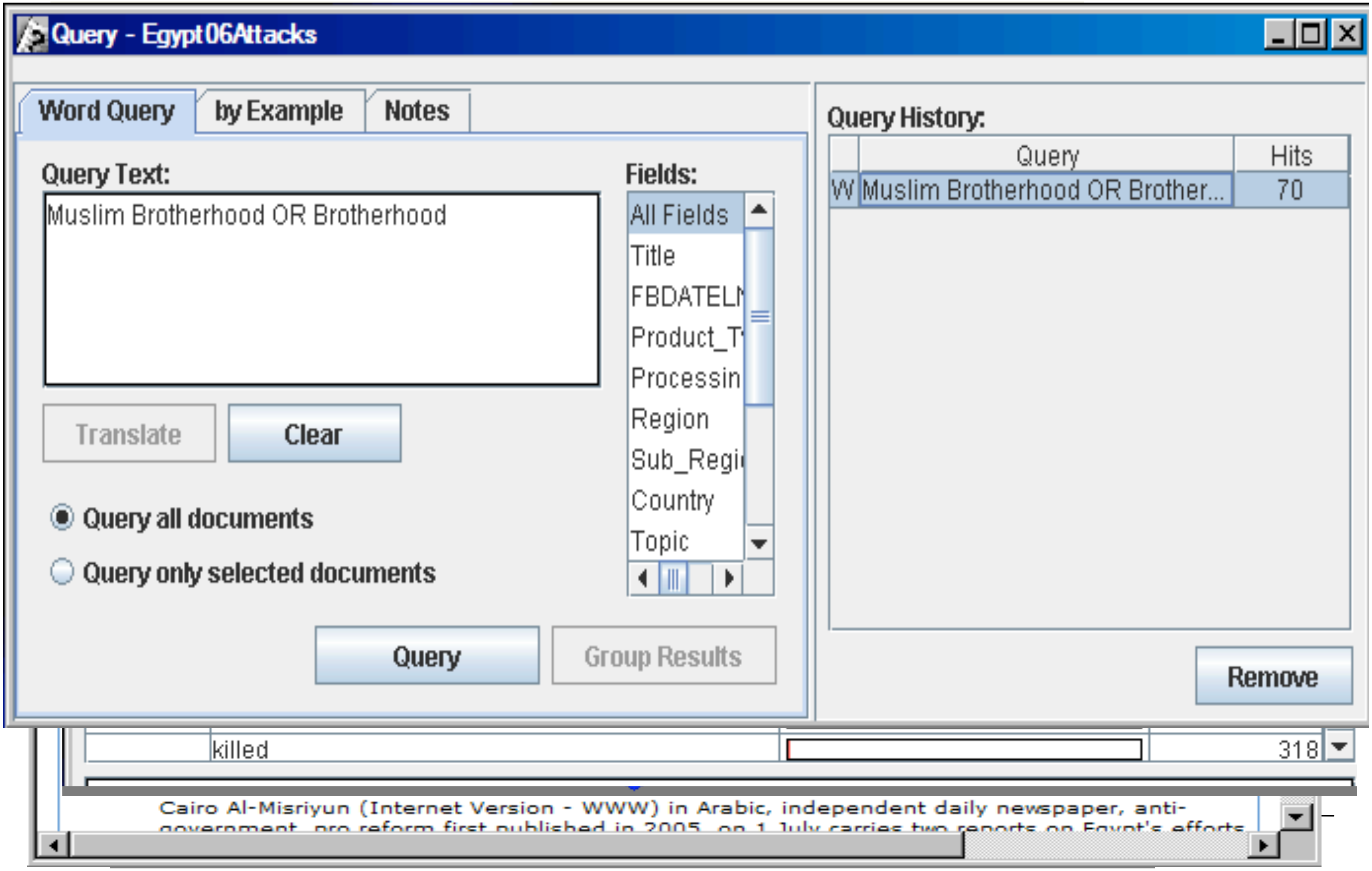
Case Study: Methodology

- Selected 1000 documents about recent terrorist attacks in Egypt
- Developed scheme of 53 codes with several hundred code keywords using insights from Social Movement Theory
- Assessed coding scheme using the Cohen kappa test to measure inter-coder agreement
- Hand-annotated 36 documents
- Formulated searches to identify and quantify frame signatures
- Utilizing the CAQDAS tool Qualrus
- Testing inter-annotator agreement
  - Three people coded the same text (931 words)
  - Used Cohen Kappa test to measure agreement in the assignment of codes to the same segments (Target Kappa Coefficient ≥ 0.7)

	Cohen Kappa Test Results
Coders A & B	0.78
Coders A & C	0.69
Coders B & C	0.73

PNNL researchers: Antonio Sanfilippo, Andrew Cowell, Annie Boek and Stephen Tratz

Case Study: Harvest, classify and select documents



Query Text: Muslim Brotherhood OR Brotherhood

● Query all documents

○ Query only selected documents

Query History:

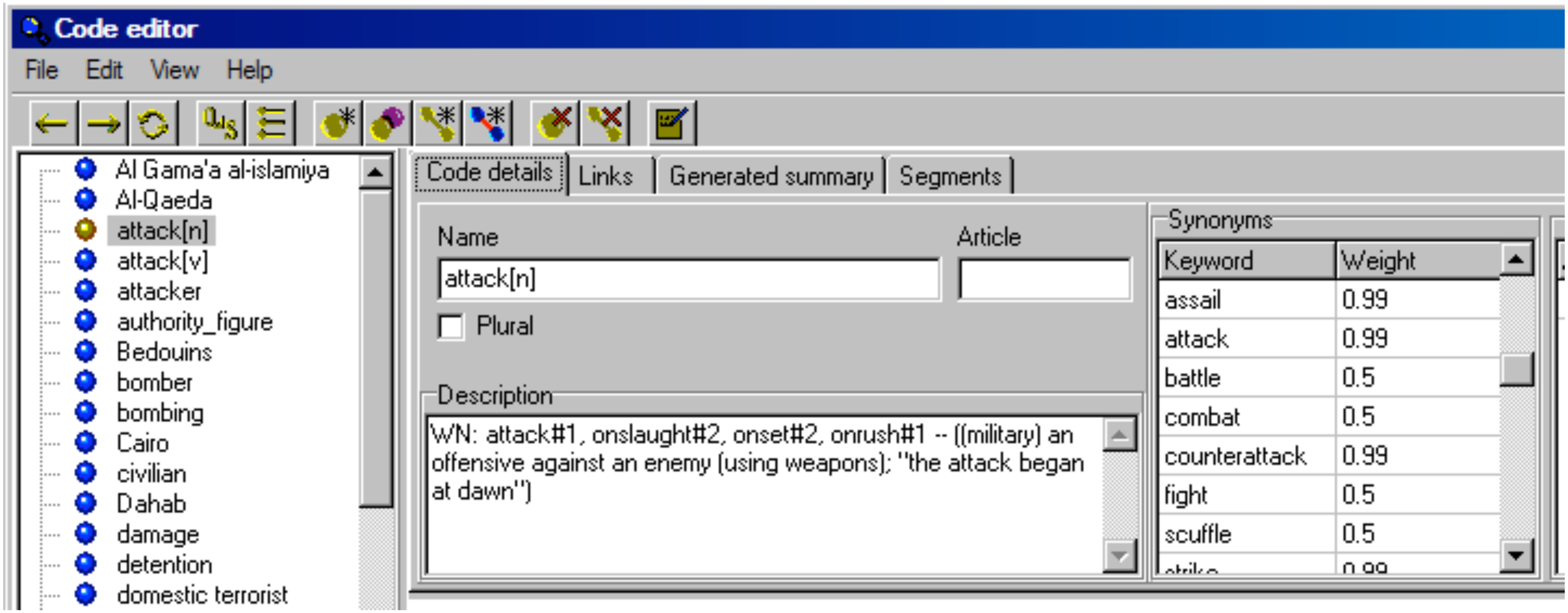
	Query	Hits
W	Muslim Brotherhood OR Brother...	70

killed 318

Cairo Al-Misriyun (Internet Version - WWW) in Arabic, independent daily newspaper, anti-government, pro reform first published in 2005, on 1 July carries two reports on Egypt's efforts

Case Study: Create coding guidelines

- Each code is associated with keywords (concepts in WordNet lexical semantic graph)
- Weights specify the similarity of the keyword to the code



- Al Gama'a al-islamiya
- Al-Qaeda
- attack[n]
- attack[v]
- attacker
- authority_figure
- Bedouins
- bomber
- bombing
- Cairo
- civilian
- Dahab
- damage
- detention
- domestic terrorist

Name: attack[n]

Article:

□ Plural

Description

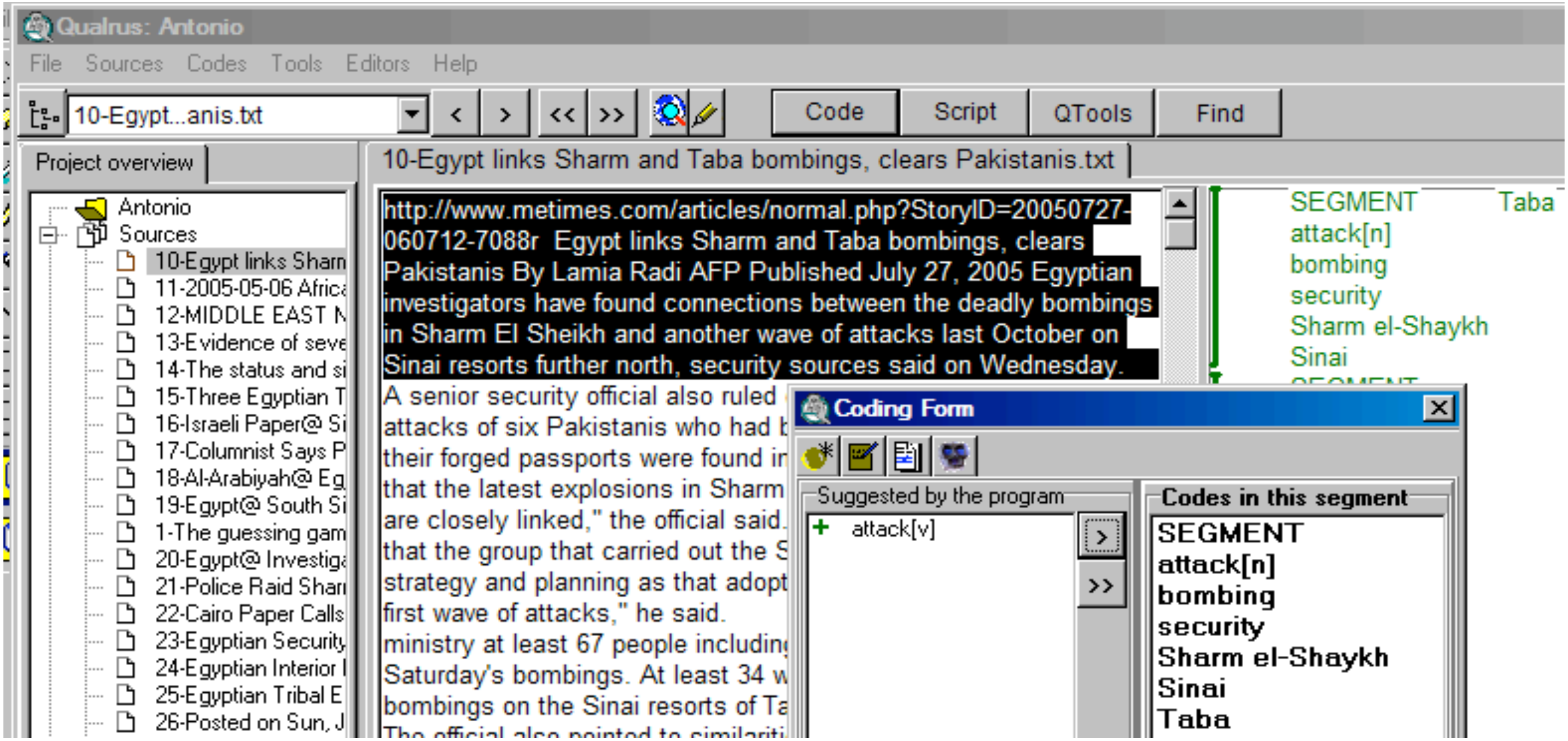
WN: attack#1, onslaught#2, onset#2, onrush#1 -- ((military) an offensive against an enemy (using weapons); "the attack began at dawn")

Synonyms

Keyword	Weight
assail	0.99
attack	0.99
battle	0.5
combat	0.5
counterattack	0.99
fight	0.5
scuffle	0.5
strike	0.99

Case Study: Annotate Documents

- Named Entity Recognition helps identify automatically segments to be coded
- Code keywords and weights derived from lexical management aid manual annotation



Antonio
Sources
10-Egypt links Sharn
11-2005-05-06 Africa
12-MIDDLE EAST N
13-Evidence of seve
14-The status and si
15-Three Egyptian T
16-Israeli Paper@ Si
17-Columnist Says P
18-Al-Arabiyah@ Eg
19-Egypt@ South Si
1-The guessing gam
20-Egypt@ Investiga
21-Police Raid Shari
22-Cairo Paper Calls
23-Egyptian Security
24-Egyptian Interior I
25-Egyptian Tribal E
26-Posted on Sun, J

10-Egypt links Sharm and Taba bombings, clears Pakistanis.txt

http://www.metimes.com/articles/normal.php?StoryID=20050727-060712-7088r Egypt links Sharm and Taba bombings, clears Pakistanis By Lamia Radi AFP Published July 27, 2005 Egyptian investigators have found connections between the deadly bombings in Sharm El Sheikh and another wave of attacks last October on Sinai resorts further north, security sources said on Wednesday.

A senior security official also ruled
attacks of six Pakistanis who had b
their forged passports were found ir
that the latest explosions in Sharm
are closely linked," the official said.
that the group that carried out the S
strategy and planning as that adopt
first wave of attacks," he said.
ministry at least 67 people includin
Saturday's bombings. At least 34 w
bombings on the Sinai resorts of Ta

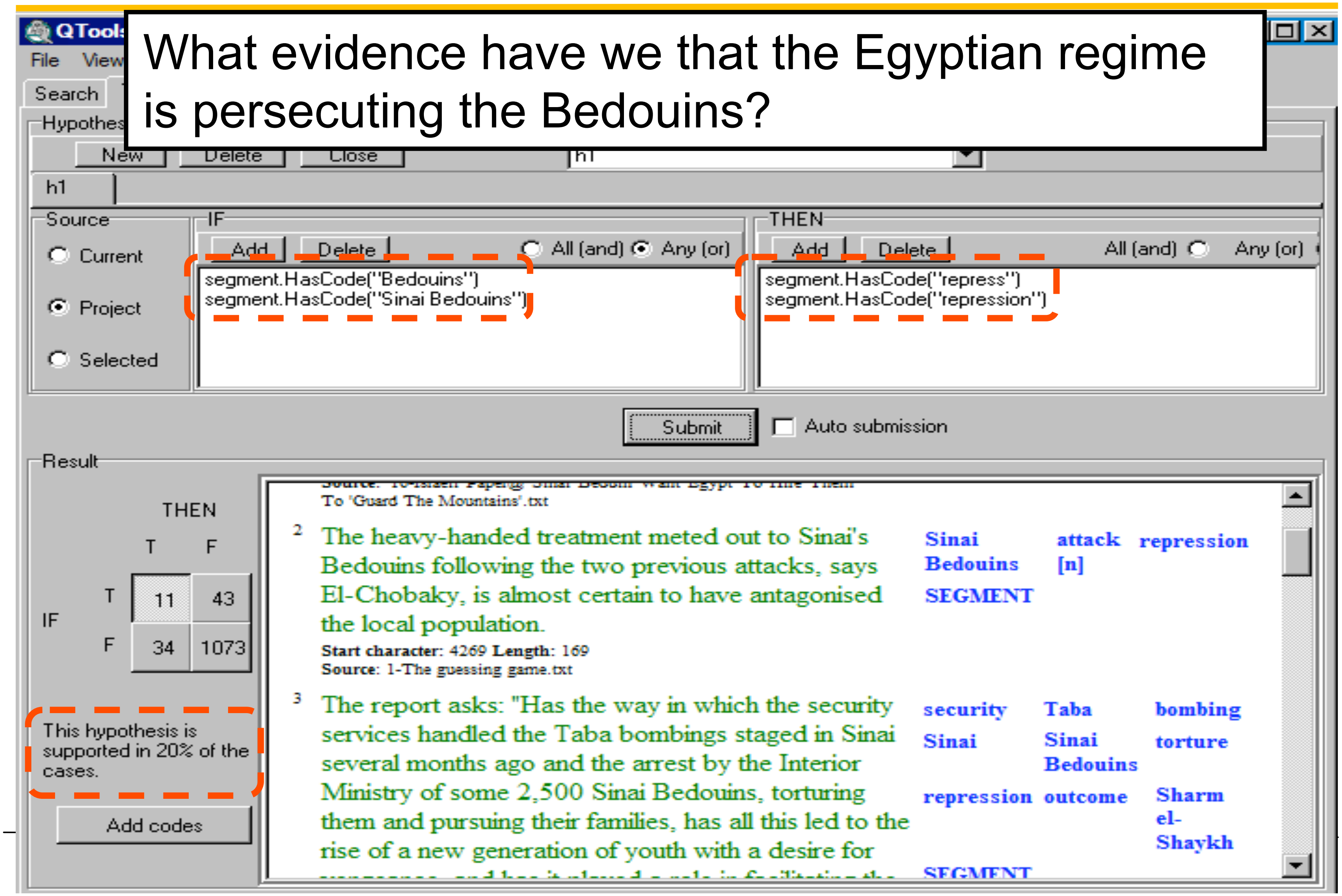
SEGMENT Taba
attack[n]
bombing
security
Sharm el-Shaykh
Sinai

Suggested by the program:
+ attack[v]

Codes in this segment:
SEGMENT
attack[n]
bombing
security
Sharm el-Shaykh
Sinai
Taba

Case Study: Construct queries

What evidence have we that the Egyptian regime is persecuting the Bedouins?



Source: ○ Current ● Project ○ Selected

IF: ○ All (and) ● Any (or)

```
segment.HasCode("Bedouins")
segment.HasCode("Sinai Bedouins")
```

THEN: All (and) ○ Any (or)

```
segment.HasCode("repress")
segment.HasCode("repression")
```

□ Auto submission

		THEN	THEN
		T	F
IF	T	11	43
IF	F	34	1073

This hypothesis is supported in 20% of the cases.

Source: 10-Israeli Paper: Sinai Bedouin Want Egypt To Hire Them To 'Guard The Mountains'.txt

2 The heavy-handed treatment meted out to Sinai's Bedouins following the two previous attacks, says El-Chobaky, is almost certain to have antagonised the local population.

Sinai Bedouins | attack [n] | repression | SEGMENT

Start character: 4269 **Length**: 169
Source: 1-The guessing game.txt

3 The report asks: "Has the way in which the security services handled the Taba bombings staged in Sinai several months ago and the arrest by the Interior Ministry of some 2,500 Sinai Bedouins, torturing them and pursuing their families, has all this led to the rise of a new generation of youth with a desire for vengeance, and has it played a role in facilitating the

security | Taba | bombing | Sinai | Sinai Bedouins | torture | repression | outcome | Sharm el-Shaykh | SEGMENT

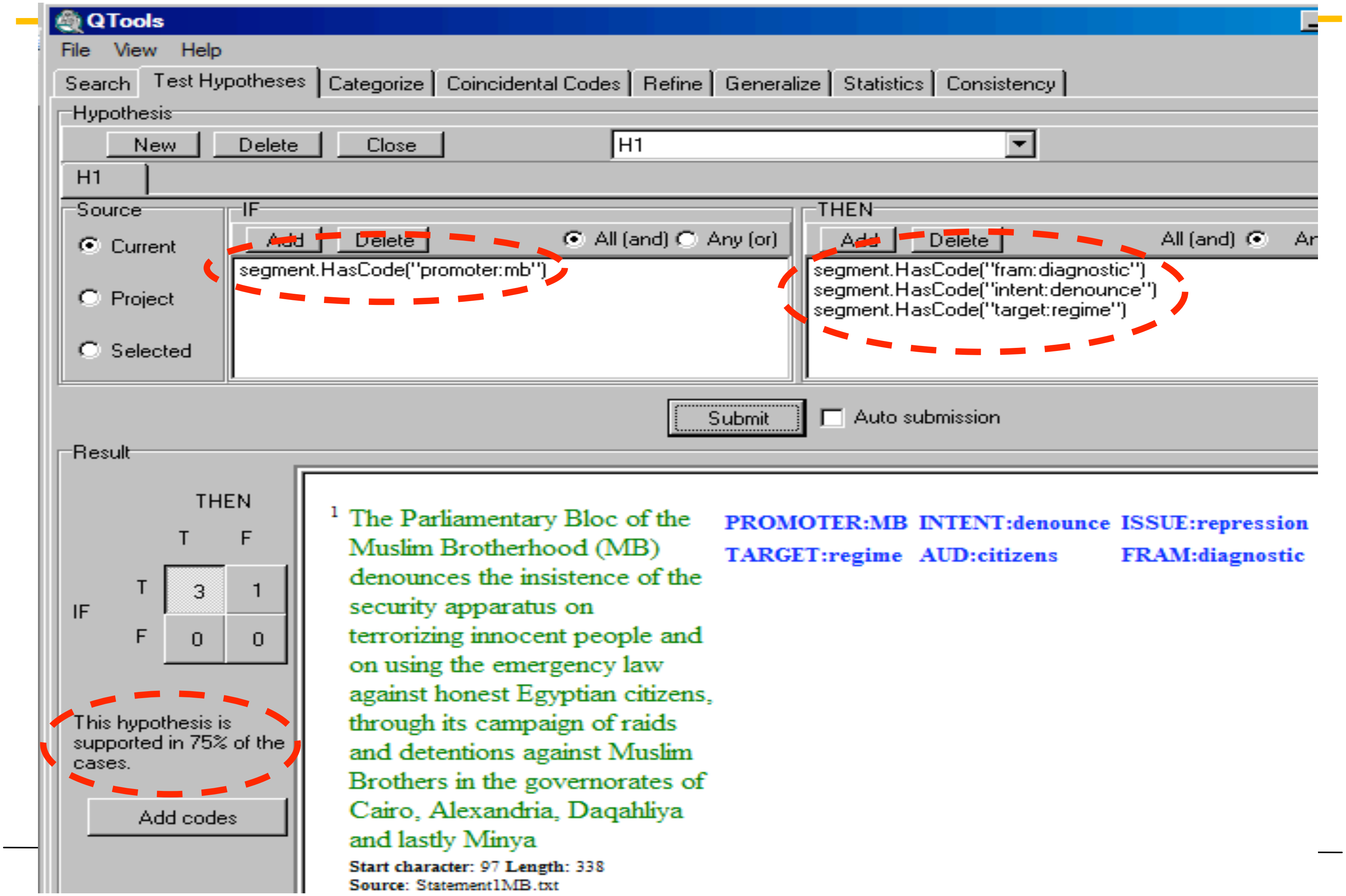
Case Study: Frame Annotation

The Parliamentary Bloc of the Muslim Brotherhood (MB) denounces the insistence of the security apparatus on terrorizing innocent people and on using the emergency law against honest Egyptian citizens, through its campaign of raids and detentions against Muslim Brothers in the governorates of Cairo, Alexandria, Daqahliya and lastly Minya.	PROMOTER
	INTENTION OBJECTIVE
	TARGET (the person or organization to blame for grievances)
	ISSUE
	AUDIENCE (the group at which the message is aimed)

- **Diagnostic anti-system Frame**
  - **Promoter** = Muslim Brotherhood
  - **Intent** = Denounce
  - **Issue** = Repression
  - **Target** = Egyptian Regime
  - **Audience** = Egyptian Citizens
- **Strength =** Supported by 75% of the data consulted
- **Provenance** = [list document sources]

- Goal: annotate larger constructs of Social Movement Theory, specifically *Frames*
- Frame constituents
  - **promoter:** *who is behind the message?*
  - **target:** *who is to blame for grievances?*
  - **intention:** *what is the objective of the message?*
  - **issue:** *what are the grievances?*
  - **audience**: *aggrieved group*
- Reason/generalize over language to establish constituent codes (e.g. *issue: repression*)

Case Study: Assess Frame Evidence



H1

Source: ● Current ○ Project ○ Selected

IF: ● All (and) ○ Any (or)

```
segment.HasCode("promoter:mb")
```

THEN: All (and) ● Ar

```
segment.HasCode("fram:diagnostic")
segment.HasCode("intent:denounce")
segment.HasCode("target:regime")
```

□ Auto submission

		THEN T	THEN F
IF	T	3	1
IF	F	0	0

This hypothesis is supported in 75% of the cases.

[1] The Parliamentary Bloc of the Muslim Brotherhood (MB) denounces the insistence of the security apparatus on terrorizing innocent people and on using the emergency law against honest Egyptian citizens, through its campaign of raids and detentions against Muslim Brothers in the governorates of Cairo, Alexandria, Daqahliya and lastly Minya

PROMOTER:MB	INTENT:denounce	ISSUE:repression
TARGET:regime	AUD:citizens	FRAM:diagnostic

Start character: 97 **Length:** 338
Source: Statement1MB.txt

Conclusions

- NLP techniques dovetail with existing methods in content analysis
- By augmenting CAQDAS tools with NLP, coding can proceed more rapidly
  - automation of basic coding
  - sophisticated tools for handling term identification and variations
  - yet, the human is still in the loop for verification
- Construction of a repository of coded texts for a specific domain can be drawn on to enable learning for further coding automation